GPQA: A Graduate-Level Google-Proof Q&A Benchmark

Year: 2023 | Citations: 1528 | Authors: David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang

Abstract

We present GPQA, a challenging dataset of 448 multiple-choice questions written by domain experts in biology, physics, and chemistry. We ensure that the questions are high-quality and extremely difficult: experts who have or are pursuing PhDs in the corresponding domains reach 65% accuracy (74% when discounting clear mistakes the experts identified in retrospect), while highly skilled non-expert validators only reach 34% accuracy, despite spending on average over 30 minutes with unrestricted access to the web (i.e., the questions are"Google-proof"). The questions are also difficult for state-of-the-art AI systems, with our strongest GPT-4 based baseline achieving 39% accuracy. If we are to use future AI systems to help us answer very hard questions, for example, when developing new scientific knowledge, we need to develop scalable oversight methods that enable humans to supervise their outputs, which may be difficult even if the supervisors are themselves skilled and knowledgeable. The difficulty of GPQA both for skilled non-experts and frontier AI systems should enable realistic scalable oversight experiments, which we hope can help devise ways for human experts to reliably get truthful information from AI systems that surpass human capabilities.